Sound source localization and speech enhancement with sparse Bayesian learning beamforming

Xenaki, Angeliki; Boldt, Jesper Bünsow; Christensen, Mads Græsbøll

Published in:
The Journal of the Acoustical Society of America

DOI (link to publication from Publisher):
10.1121/1.5042222

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Sound source localization and speech enhancement with sparse Bayesian learning beamforming

Angeliki Xenaki[a] and Jesper Bünsow Boldt
GN Hearing A/S, DK-2750 Ballerup, Denmark

Mads Græsbøll Christensen
Audio Analysis Lab, AD:MT, Aalborg University, DK-9000 Aalborg, Denmark

(Received 27 October 2017; revised 1 February 2018; accepted 22 February 2018; published online 29 June 2018)

Speech localization and enhancement involves sound source mapping and reconstruction from noisy recordings of speech mixtures with microphone arrays. Conventional beamforming methods suffer from low resolution, especially with a limited number of microphones. In practice, there are only a few sources compared to the possible directions-of-arrival (DOA). Hence, DOA estimation is formulated as a sparse signal reconstruction problem and solved with sparse Bayesian learning (SBL). SBL uses a hierarchical two-level Bayesian inference to reconstruct sparse estimates from a small set of observations. The first level derives the posterior probability of the complex source amplitudes from the data likelihood and the prior. The second level tunes the prior towards sparse solutions with hyperparameters which maximize the evidence, i.e., the data probability. The adaptive learning of the hyperparameters from the data auto-regularizes the inference problem towards sparse robust estimates. Simulations and experimental data demonstrate that SBL beamforming provides high-resolution DOA maps outperforming traditional methods especially for correlated or non-stationary signals. Specifically for speech signals, the high-resolution SBL reconstruction offers not only speech enhancement but effectively speech separation.
 https://doi.org/10.1121/1.5042222

[WS] Pages: 3912–3921

I. INTRODUCTION

Talker localization and separation are key aspects in computational auditory scene analysis, i.e., the segregation of sources from noisy and reverberant sound mixtures with signal processing. Multi-microphone processing systems are able to exploit both the spatial and spectral information of the wavefield thus have improved performance compared to single-microphone systems.[1,2] Multi-channel speech localization and enhancement algorithms find several applications including robot audition,[3,4] tele-conferencing,[5] and hearing aids.[6,7]

The problem of sound source localization in array signal processing is to infer the direction-of-arrival (DOA) of the source signals from noisy measurements of the wavefield with an array of microphones. Beamforming methods based on spatial filtering have low resolution or degraded performance for coherent arrivals, e.g., in reverberant conditions, or for non-stationary signals, when only a few observation windows (snapshots) are available.[8] In acoustic imaging, there are usually only a few sources generating the observed wavefield such that the DOA map is sparse, i.e., it can be fully described by only a few parameters. Exploiting the underlying sparsity, sparse signal reconstruction improves significantly the resolution in DOA estimation.[9–12] While ℓ_p-norm regularized maximum likelihood methods, with $p \leq 1$, have been proposed to promote sparsity in DOA estimation[9–11,13] and wavefield reconstruction,[14,15] the accuracy of the resulting sparse estimate is determined by the *ad hoc* choice of the regularization parameter.[12,16]

Sparse Bayesian learning (SBL) is a probabilistic parameter estimation approach which is based on a hierarchical Bayesian method for learning sparse models from possibly overcomplete representations resulting in robust maximum likelihood estimates.[17,18] Specifically, the Bayesian formulation of SBL allows regularizing the maximum likelihood estimate with prior information on the model parameters. However, instead of explicitly introducing specialized model priors to reflect the underlying structure, SBL uses a hierarchical model which controls the scaling of a multivariate Gaussian prior distribution through individual hyperparameters for each model parameter. The hyperparameters are iteratively estimated from the data selecting the most relevant model features while practically nulling the probability of irrelevant features, hence promoting sparsity.[17,19] Since SBL *learns* the hyperparameters from the data, it allows for *automatic* regularization of the maximum likelihood estimate which adapts to the problem under study.[17,20] The hierarchical formulation of SBL inference offers both a computationally convenient Gaussian posterior distribution for adaptive processing (type-I maximum likelihood) and automatic regularization towards robust sparse estimates determined by the hyperparameters which maximize the evidence (type-II maximum likelihood).[21]

In array signal processing, SBL is shown to improve significantly the resolution in beamforming[22] and in general the accuracy of DOA estimation,[23–28] outperforming conventional

[a]Electronic mail: axenaki@gnresound.com

methods notably at demanding scenarios with correlated or non-stationary signals. Multi-snapshot[26] and multi-frequency[23,24,27,28] SBL inference exploits the common sparsity profile across snapshots for stationary signals and frequencies for broadband signals to provide robust estimates by alleviating the ambiguity in the spatial mapping between sources and sensors due to noise and frequency-dependent spatial aliasing, respectively. Accounting for the statistics of modelling errors in SBL estimation, e.g., due to sensor position, sound speed uncertainty or basis mismatch, further improves support recovery.[28,29]

We use the SBL framework to solve the sound source localization problem of speech mixtures in noisy and reverberant conditions. We employ the multi-snapshot, multi-frequency SBL algorithm in Ref. 28 to reconstruct simultaneously the DOA and the complex amplitude of speech signals. SBL beamforming assumes a predefined spatial mapping between the sources and the sensors to infer the DOAs directly from the reconstructed source vector, as opposed to methods (including SBL-based[7]) which infer the DOA of a single target talker indirectly through the estimation of the relative transfer function between a pair of microphones.[4,6] It is demonstrated both with simulations and experimental data that SBL beamforming offers unambiguous source localization outperforming traditional beamforming methods especially for correlated signals and single-snapshot measurements. The high-resolution SBL reconstruction offers not only speech enhancement over noise, but also speech separation between competing talkers.

Herein, vectors and matrices are represented by bold lowercase and uppercase letters, respectively. The superscripts ^{T} and ^{H} denote the transpose and the Hermitian, i.e., conjugate transpose, operator, respectively, on vectors and matrices. The superscript $^+$ denotes the generalized inverse operator on a matrix. A $Q \times Q$ identity matrix is denoted $\mathbf{I}_Q$. The ℓ_p-norm of a vector $\mathbf{x} \in \mathbb{C}^Q$ is defined as $\|\mathbf{x}\|_p = (\sum_{q=1}^{Q} |x_q|^p)^{1/p}$. The Frobenius norm of a matrix $\mathbf{X} \in \mathbb{C}^{Q \times R}$ is defined as $\|\mathbf{X}\|_F = (\sum_{q=1}^{Q} \sum_{r=1}^{R} |x_{qr}|^2)^{1/2}$.

II. ARRAY SIGNAL MODEL

Assuming narrowband processing, the complex-valued measurements at an M-element array, i.e., the data, are described by the vector

$$\mathbf{y}_l = [y_1(f,l), \ldots, y_M(f,l)]^T, \tag{1}$$

where $y_m(f, l)$ is the short-time Fourier transform (STFT) coefficient for the fth frequency and the lth time-frame (snapshot) of the recorded signal at the mth sensor, $m \in \{1, \ldots, M\}$. The frequency index f is omitted from the vector's notation for simplicity.

At the far-field of the array, the location of a source is characterized by the DOA, θ, of the associated plane wave, x. Discretizing the angular space of interest into N directions, the vector of the complex-valued sound source amplitudes, i.e., the model parameters, for the fth frequency and the lth snapshot is

$$\mathbf{x}_l = [x_1(f,l), \ldots, x_N(f,l)]^T. \tag{2}$$

The array measurements are related to the model parameters with the linear model

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{N}, \tag{3}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L] \in \mathbb{C}^{M \times L}$ is the wavefield measurements at M sensors for L snapshots, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L] \in \mathbb{C}^{N \times L}$ is the unknown source amplitudes at N angular directions for L snapshots and $\mathbf{N} \in \mathbb{C}^{M \times L}$ is additive noise which is assumed independent across sensors and snapshots. The sensing matrix,

$$\mathbf{A} = [\mathbf{a}(\theta_1), \ldots, \mathbf{a}(\theta_N)], \tag{4}$$

has as columns the steering vectors $\mathbf{a}(\theta_n)$ at each direction θ_n, $n \in \{1, \ldots, N\}$, which describe the acoustic transfer function from a source at θ_n to all M sensors on the array. The sensing matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ is determined either analytically for simple array geometries, e.g., uniform linear arrays (ULA),[11] spherical arrays baffled on a rigid sphere,[30] or experimentally, e.g., from head-related transfer function (HRTF) measurements.[31]

III. DOA ESTIMATION

The problem of DOA estimation and source reconstruction with sensor arrays[32] is to recover the sources $\mathbf{X}$, given the sensing matrix $\mathbf{A}$ and a set of observations $\mathbf{Y}$. Usually, there are only a few sources $K \ll N$ generating the acoustic field such that $\mathbf{X}$ is sparse in the angular space, i.e., has only a few non-zero components. However, precise localization requires fine angular resolution such that $M < N$ and the problem in Eq. (3) is underdetermined, i.e., has infinitely many solutions.

An estimate $\widehat{\mathbf{X}}$ can be obtained by spatial filtering the array data $\mathbf{Y}$ (beamforming), or by solving Eq. (3) with optimization or probabilistic methods for parameter estimation. For stationary sources, when $\mathbf{X}$ has a common row-wise sparsity profile, snapshots can be combined to improve the signal-to-noise ratio. Otherwise, the problem should be solved independently for each snapshot.

A. Spatial filtering

Spatial filtering of the recorded wavefield refers to applying direction-dependent complex weights $\mathbf{w}(\theta) \in \mathbb{C}^{M \times 1}$ to the sensor outputs to allow signals from a specific look-direction to pass undistorted while attenuating wavefield contributions from other directions. Applying a set of spatial weights, one for each look-direction to steer the beamformer across the angular space yields the DOA estimate,

$$\widehat{\mathbf{X}}_{\mathrm{BF}} = \mathbf{W}^H \mathbf{Y}, \tag{5}$$

where $\mathbf{W} = [\mathbf{w}(\theta_1), \ldots, \mathbf{w}(\theta_N)]$ has as columns the spatial weight vectors at each DOA θ_n, $n \in \{1, \ldots, N\}$.

Accordingly, the beamformer power at direction θ is

$$P_{\mathrm{BF}}(\theta) = \mathbf{w}^H(\theta)\mathbf{S}_{\mathbf{y}}\mathbf{w}(\theta), \tag{6}$$

where $\mathbf{S}_{\mathbf{y}} = (\mathbf{Y}\mathbf{Y}^H)/L$ is the sample data cross-spectral matrix from L snapshots. Note that, for broadband signals, spatial filtering methods are applied to each frequency separately according to the narrowband signal model Eq. (3).

1. Conventional beamforming

The conventional beamforming (CBF) is the simplest source localization method. The method uses the steering vectors as spatial weights, i.e.,

$$\mathbf{w}_{\mathrm{CBF}}(\theta) = \frac{1}{M}\mathbf{a}(\theta), \tag{7}$$

to combine the sensor outputs coherently enhancing the signal at the look-direction from the ubiquitous noise. CBF is robust to noise and can be used even with single snapshot data, $L = 1$, but is characterized by low resolution and the presence of sidelobes.

2. Minimum variance distortionless response beamforming

The minimum variance distortionless response (MVDR) beamforming[33] weight vector is obtained by minimizing the output power of the beamformer under the constraint that the signal from the look direction, θ, remains undistorted,

$$\min_{\mathbf{w}} \mathbf{w}^H\mathbf{S}_{\mathbf{y}}\mathbf{w} \text{ subject to } \mathbf{w}^H\mathbf{a}(\theta) = 1, \tag{8}$$

resulting in the optimal weight vector,

$$\mathbf{w}_{\mathrm{MVDR}}(\theta) = \frac{\left(\mathbf{S}_{\mathbf{y}} + \beta\mathbf{I}_M\right)^{-1}\mathbf{a}(\theta)}{\mathbf{a}(\theta)^H\left(\mathbf{S}_{\mathbf{y}} + \beta\mathbf{I}_M\right)^{-1}\mathbf{a}(\theta)}, \tag{9}$$

where diagonal loading with regularization parameter β is used to regularize the inverse of the sample covariance matrix $\mathbf{S}_{\mathbf{y}}^{-1}$ whenever it is rank deficient. Note that by replacing the data sample covariance matrix $\mathbf{S}_{\mathbf{y}}$ with the noise sample covariance matrix $\mathbf{S}_{\mathbf{n}} = (\mathbf{N}\mathbf{N}^H)/L$ in Eqs. (8) and (9) results in an equivalent derivation of the MVDR weights.[32] However, in practical applications it is more difficult to obtain a robust estimate of the noise separately from the measured data. MVDR beamforming offers high resolution DOA maps but its performance degrades significantly under snapshot-starved data, $L < M$, correlated arrivals and low SNR conditions.

B. Probabilistic parameter estimation

The problem of DOA estimation can be formulated in a probabilistic framework by considering both the unknowns $\mathbf{X}$ and the observations $\mathbf{Y}$ as stochastic processes and solved with Bayesian inference.[16]

Bayes' theorem,

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})}, \tag{10}$$

derives the posterior distribution $p(\mathbf{X}|\mathbf{Y})$ of the model parameters $\mathbf{X}$, i.e., the complex source amplitudes, conditioned on the data $\mathbf{Y}$, i.e., the sensor measurements, from the data likelihood $p(\mathbf{Y}|\mathbf{X})$, the prior distribution of the model parameters $p(\mathbf{X})$ and the marginal distribution of the data $p(\mathbf{Y})$. The maximum *a posteriori* (MAP) estimate,

$$\begin{aligned}\widehat{\mathbf{X}}_{\mathrm{MAP}} &= \arg\max_{\mathbf{X}} \ln p(\mathbf{X}|\mathbf{Y}) \\ &= \arg\min_{\mathbf{X}} \left[-\ln p(\mathbf{Y}|\mathbf{X}) - \ln p(\mathbf{X})\right],\end{aligned} \tag{11}$$

is used for DOA reconstruction. Here, $p(\mathbf{Y})$ is omitted from the optimization as it is marginalized over $\mathbf{X}$.

The probabilistic formulation (11) provides a regularized solution to the DOA estimation problem (3) based on prior information. To demonstrate the effect of prior information on the estimate, consider the single-snapshot case. Assuming that the additive noise is independent and identically distributed (iid) circularly symmetric complex Gaussian with variance σ^2, $p(\mathbf{n};\sigma^2) = \mathcal{CN}(\mathbf{n}|0,\sigma^2\mathbf{I})$, the data likelihood is also complex Gaussian distributed,

$$p(\mathbf{y}|\mathbf{x};\sigma^2) = \mathcal{CN}(\mathbf{y}|\mathbf{A}\mathbf{x},\sigma^2\mathbf{I}) \propto \mathrm{e}^{-(\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2/\sigma^2)}. \tag{12}$$

Employing a general expression for the prior $p(\mathbf{x})$ based on the multivariate generalized complex Gaussian distribution,[34]

$$p(\mathbf{x};\nu^p) \propto \mathrm{e}^{-(\|\mathbf{x}\|_p/\nu)^p}, \tag{13}$$

where $\nu \in \mathbb{R}_+$ is the scaling parameter and $p \in \mathbb{R}_+$ is the shape parameter, the MAP estimate (11) is expressed as a regularized least-squares (R-LS) problem,

$$\widehat{\mathbf{x}}_{\mathrm{R\text{-}LS}}(p,\mu) = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_p^p, \tag{14}$$

where $\mu = \sigma^2/\nu^p \geq 0$ is the regularization parameter which controls the relative importance between the data fit and the regularization term. The characteristics of the MAP estimate depend on the choice of the shape parameter p and the regularization parameter μ.[12]

For example, assuming that the model parameters follow an iid complex Gaussian distribution, $p(\mathbf{x};\nu^2) = \mathcal{CN}(\mathbf{x}|0,\nu^2\mathbf{I})$, problem (14) becomes an ℓ_2-norm regularized least-squares problem which has an analytic solution,

$$\begin{aligned}\widehat{\mathbf{x}}_{\ell_2}(\mu) &= \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_2^2 \\ &= \mathbf{A}^H(\mathbf{A}\mathbf{A}^H + \mu\mathbf{I}_M)^{-1}\mathbf{y}.\end{aligned} \tag{15}$$

The ℓ_2-norm regularizer penalizes the energy in the solution hence the estimate (15) is smooth and robust to noise but has low resolution. Note that CBF is related to the ℓ_2-norm estimate for large μ,[12]

$$\widehat{\mathbf{x}}_{\mathrm{CBF}} = \lim_{\mu\to\infty}\left[\mu\widehat{\mathbf{x}}_{\ell_2}(\mu)\right] = \mathbf{A}^H\mathbf{y}. \tag{16}$$

Contrarily, assuming that the model coefficients follow a Laplacian-like distribution for complex random variables,[35]

$$p(\mathbf{x};\nu) \propto \mathrm{e}^{-(\|\mathbf{x}\|_1/\nu)}, \tag{17}$$

the MAP estimate (14) becomes the solution to an ℓ_1-norm regularized least-squares problem,

$$\widehat{\mathbf{x}}_{\ell_1}(\mu) = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\|x\|_1, \tag{18}$$

which is known as the least absolute shrinkage and selection operator[36] (Lasso) since the ℓ_1-norm regularizer shrinks the model coefficients towards zero as the regularization parameter, $\mu = \sigma^2/\nu$, increases.

As opposed to a Gaussian prior, the Laplacian-like prior distribution encourages sparse solutions as it concentrates more mass at zero and in the tails. Thus the ℓ_1-norm estimate improves significantly the resolution in DOA estimation in the presence of only a few sources.[11,12] The ℓ_1-norm minimization problem (18) can be solved with convex optimization algorithms[37] which can be computationally intensive. Besides, the accuracy of the ℓ_1-norm estimate (18) depends on the regularization parameter which determines the degree of sparsity in the estimate and requires knowledge on the hyperparameters, i.e., σ^2 and ν, of the underlying probability distributions (12) and (17).

1. Sparse Bayesian learning beamforming

The SBL framework uses a hierarchical approach to probabilistic parameter estimation. Instead of employing specialized prior models, e.g., Eq. (17), to explicitly promote sparse maximum likelihood estimates, e.g., Eq. (18), SBL uses a Gaussian prior, $p(\mathbf{x};\boldsymbol{\xi}) = \mathcal{CN}(\mathbf{x}|0,\Xi)$, with diagonal covariance matrix $\Xi = \mathrm{diag}(\boldsymbol{\xi})$ and controls the sparsity in the estimate by scaling the model parameters, $\mathbf{x}$, with individual hyperparameters, $\boldsymbol{\xi}$. The hyperparameters $\boldsymbol{\xi}$ are estimated from the data and control the variances of each coefficient in $\mathbf{x}$, i.e., the source powers. Given that the model parameters are independent across snapshots, the multi-snapshot prior distribution is

$$p(\mathbf{X}) = \prod_{l=1}^{L} \mathcal{CN}(\mathbf{x}_l|0,\Xi). \tag{19}$$

Similarly, assuming that the noise is zero-mean complex Gaussian, independent both across sensors and snapshots such that $p(\mathbf{N}) = \prod_{l=1}^{L} \mathcal{CN}(\mathbf{n}_l|0,\boldsymbol{\Sigma}_\mathbf{n})$ with covariance matrix $\boldsymbol{\Sigma}_\mathbf{n} = \sigma^2\mathbf{I}$, the multi-snapshot data likelihood is

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{l=1}^{L} \mathcal{CN}(\mathbf{y}_l|\mathbf{A}\mathbf{x}_l,\sigma^2\mathbf{I}). \tag{20}$$

Given the Gaussian prior (19) and likelihood (20) for independent snapshots, the posterior distribution for $\mathbf{X}$ is also Gaussian,

$$p(\mathbf{X}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}) = \prod_{l=1}^{L} \mathcal{CN}(\mathbf{m}_l,\Sigma_\mathbf{x}), \tag{21}$$

where

$$\mathbf{m}_l = \Xi\mathbf{A}^H\Sigma_\mathbf{y}^{-1}\mathbf{y}_l, \quad l \in \{1,\ldots,L\}, \tag{22}$$

$$\Sigma_\mathbf{x} = \Xi - \Xi\mathbf{A}^H\Sigma_\mathbf{y}^{-1}\mathbf{A}\Xi \tag{23}$$

is the posterior mean and covariance, respectively, and

$$\Sigma_\mathbf{y} = E\{\mathbf{y}_l\mathbf{y}_l^H\} = \mathbf{A}\Xi\mathbf{A}^H + \sigma^2\mathbf{I} \tag{24}$$

is the data covariance matrix. Given the hyperparameters Ξ, or simply $\boldsymbol{\xi}$ since Ξ is considered diagonal, and σ^2, the MAP estimate (11) is the posterior mean (22), $\widehat{\mathbf{X}}_{\mathrm{MAP}}(\boldsymbol{\xi},\sigma^2) = [\mathbf{m}_1,\ldots,\mathbf{m}_L]$. Note that the sparsity of $\widehat{\mathbf{X}}_{\mathrm{MAP}}$ is dictated by the sparsity profile of the hyperparameters $\boldsymbol{\xi}$, i.e., $x_n = 0$ if $\xi_n = 0$, $n \in \{1,\ldots,N\}$.

In SBL the hyperparameters $\boldsymbol{\xi}$ and σ^2 are estimated from the evidence, i.e., the unconditional probability distribution of the data marginalized over the model parameters $\mathbf{X}$,

$$\begin{aligned} p(\mathbf{Y}) &= \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X} \\ &= \int \prod_{l=1}^{L} \mathcal{CN}(\mathbf{y}_l|\mathbf{A}\mathbf{x}_l,\sigma^2\mathbf{I})\mathcal{CN}(\mathbf{x}_l|0,\Xi)\mathrm{d}\mathbf{X} \\ &= \prod_{l=1}^{L} \mathcal{CN}(\mathbf{y}_l|0,\boldsymbol{\Sigma}_\mathbf{y}). \end{aligned} \tag{25}$$

First, the hyperparameters $\widehat{\boldsymbol{\xi}}$ are estimated with a type-II maximum likelihood, i.e., by maximizing the evidence,

$$\begin{aligned} \widehat{\boldsymbol{\xi}} &= \arg\max_{\xi\geq 0} \log p(\mathbf{Y}) = \arg\max_{\xi\geq 0} \log \frac{e^{-\mathrm{Tr}\left(\mathbf{Y}^H\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{Y}\right)}}{\left(\pi^M \det\left(\boldsymbol{\Sigma}_\mathbf{y}\right)\right)^L} \\ &= \arg\min_{\xi\geq 0}\left\{L\log\det\left(\boldsymbol{\Sigma}_\mathbf{y}\right) + \mathrm{Tr}\left(\mathbf{Y}^H\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{Y}\right)\right\}, \end{aligned} \tag{26}$$

where $\mathrm{Tr}(\cdot)$ and $\det(\cdot)$ denote, respectively, the trace and determinant operators on a matrix. The objective function of the resulting minimization problem (26) is non-convex.[37] However, problem (26) can be solved approximately by differentiating the objective function to obtain the fixed point updates,[26,28]

$$\widehat{\xi}_n^i = \widehat{\xi}_n^{i-1} \frac{\mathbf{a}(\theta_n)^H\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{S}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{a}(\theta_n)}{\mathbf{a}(\theta_n)^H\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{a}(\theta_n)}, \tag{27}$$

where $\widehat{\xi}_n^i$ is the estimated variance of the nth model parameter, i.e., the estimated source power of a source at direction θ_n, at the ith iteration.

Then, the estimation of the hyperparameter σ^2 is based on a stochastic maximum likelihood procedure,[26]

$$\widehat{\sigma}^2 = \frac{1}{M-K}\mathrm{Tr}\left[\left(\mathbf{I}_M - \mathbf{A}_\mathcal{N}\mathbf{A}_\mathcal{N}^+\right)\mathbf{S}_\mathbf{y}\right], \tag{28}$$

where $\mathcal{N} = \{n \in \mathbb{N}|\mathrm{K\ largest\ peaks\ in}\ \boldsymbol{\xi}^i\} = \{n_1,\ldots,n_K\}$ is the set of the active indices indicating the position of the K largest peaks in $\widehat{\boldsymbol{\xi}}^i$ such that $\mathbf{A}_\mathcal{N} = [\mathbf{a}(\theta_{n_1}),\ldots,\mathbf{a}(\theta_{n_K})]$.

To this point, the derivation is based on the narrowband model (3). For broadband signals, we can exploit the common sparsity profile across frequencies to enhance the sparsity of the estimate $\widehat{\boldsymbol{\xi}}$. The narrowband estimates $\widehat{\boldsymbol{\xi}}(f)$ Eq. (27) can be either combined incoherently for F frequencies,

$$\widehat{\boldsymbol{\xi}}_{1:F} = \frac{1}{F}\sum_{f=1}^{F}\widehat{\boldsymbol{\xi}}(f), \tag{29}$$

or coherently assuming a prior with common covariance Ξ across frequencies, $p[\mathbf{X}(f)] = \prod_{l=1}^{L}\mathcal{CN}(\mathbf{x}_l(f)|0,\Xi)$, $\forall f \in \{1,\ldots,F\}$ which results to a unified update rule for all frequencies,[28]

$$\widehat{\xi}^{\,i}_{n,1:F} = \widehat{\xi}^{\,i-1}_{n,1:F}\frac{\sum_{f=1}^{F}\mathbf{a}(\theta_n,f)^H\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(f)\mathbf{S}_{\mathbf{y}}(f)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(f)\mathbf{a}(\theta_n,f)}{\sum_{f=1}^{F}\mathbf{a}(\theta_n,f)^H\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(f)\mathbf{a}(\theta_n,f)}. \tag{30}$$

Table I summarizes the algorithm for SBL DOA estimation. The beamformer power spectrum is readily given by the hyperparameters $\widehat{\boldsymbol{\xi}}$ which represent source power. For amplitude reconstruction, the unbiased estimate, $\widehat{\mathbf{X}}_{\mathrm{SBL},\mathcal{N}} = \mathbf{A}_{\mathcal{N}}^{+}\mathbf{Y}$, is used instead of the MAP estimate $\widehat{\mathbf{X}}_{\mathrm{MAP}}(\widehat{\boldsymbol{\xi}},\widehat{\sigma}^2) = [\mathbf{m}_1,\ldots,\mathbf{m}_L]$ as it provides more accurate estimates.[38] Nevertheless, highly correlated steering vectors, e.g., at very low frequencies, will increase the condition number of $\mathbf{A}_{\mathcal{N}}$ and, consequently, the error in the corresponding matrix inversion. For narrowband estimation set $F = 1$, in which case the update rules (29) and (30) are equivalent, i.e., they reduce to Eq. (27). The details of the derivation of the hyperparameter update rules Eqs. (27) and (28) and of the algorithm for the implementation of the SBL beamformer are in Refs. 26 and 28.

Note that the descritization of the problem (3) to a predefined angular grid may affect the accuracy of the SBL estimate. This is either due to basis mismatch for grids that are too coarse to capture the true DOAs of the signals or due to high correlation of adjacent steering vectors for dense grids. Such uncertainty can be incorporated to the model as additive or multiplicative noise and the effect of modelling error can be mitigated by tuning the hyperparameters that control its statistics.[28,29] In the interest of algorithm simplicity for practical applications, modelling errors are neglected herein.

TABLE I. Algorithm for SBL beamforming.

Inputs: $\mathbf{A}$, $\mathbf{Y}$, $\mathbf{S}_{\mathbf{y}}$, $\forall f, f \in \{1,\ldots,F\}$
Initializations: $i = 0$, $\epsilon = 1$, $\widehat{\boldsymbol{\xi}}^{\,i} = \mathbf{1}$, $\widehat{\sigma}^2 = 0.1$, $\forall f$
Parameters: N_{iter}, ϵ_{min}, K
1: **while** $i < N_{\mathrm{iter}}$ **and** $\epsilon > \epsilon_{\mathrm{min}}$
2: Update $i = i + 1$
3: Compute $\boldsymbol{\Sigma}_{\mathbf{y}}$ using (24), $\forall f$
4: Update $\widehat{\boldsymbol{\xi}}^{\,i}$ using $\begin{cases}(27),(29), & \text{or}\\ (30)\end{cases}$
5: Find $\mathcal{N} = \{n \in \mathbb{N} | \text{K largest peaks in } \widehat{\boldsymbol{\xi}}^{\,i}\}$
6: Update $\widehat{\sigma}^2$ using (28), $\forall f$
7: Update $\epsilon = \frac{\|\widehat{\boldsymbol{\xi}}^{\,i} - \widehat{\boldsymbol{\xi}}^{\,i-1}\|_1}{\|\widehat{\boldsymbol{\xi}}^{\,i-1}\|_1}$
8: **end**
Output: $\widehat{\boldsymbol{\xi}}$, $\widehat{\sigma}^2$, $\mathcal{N}$
Signal estimate: $\widehat{\mathbf{X}}_{\mathrm{SBL},\mathcal{N}} = \mathbf{A}_{\mathcal{N}}^{+}\mathbf{Y}$
Beamformer power: $P_{\mathrm{SBL}}(\theta_n) = \widehat{\xi}_n$, $n \in \{1,\ldots,N\}$

Moreover, we assume K known in step 5 of the algorithm in Table I, otherwise it can be determined with model order identification methods.[23]

C. Comparison of beamforming methods

Figure 1 compares the DOA power spectra of CBF, MVDR, and SBL beamformer [Eq. (6) and $\widehat{\boldsymbol{\xi}}$, respectively] on a simple configuration with a ULA. For a ULA with M sensors, the sensing matrix (4) is defined by the steering vectors,[32]

$$\mathbf{a}(\theta_n) = e^{j2\pi(d/\lambda)[0,\ldots,M-1]^T\sin\theta_n}, \tag{31}$$

where d is the uniform inter-sensor spacing, λ is the wavelength and θ_n is the nth direction of arrival with respect to the array axis. To demonstrate the resolution capabilities of the beamformers, two sources are introduced with equal deterministic amplitude and random phase uniformly distributed in $[0, 2\pi)$ on a grid with angular spacing 5°. Note that the DOA spectrum is limited within $[-90°, 90°]$ due to the left-right ambiguity of ULA [i.e., $\sin\theta = \sin(\pi - \theta)$].[32] The noise variance is determined by the SNR given the average source power across snapshots, $\sigma^2 = 10^{-\mathrm{SNR}/10}\|\mathbf{X}\|_F^2/L$.

For uncorrelated sources, high SNR and sufficient snapshots, all beamforming methods indicate the presence of the two sources as peaks in the power spectrum, albeit CBF with low-resolution and a prominent sidelobe at around −50°, Fig. 1(a). The high-resolution performance of the MVDR beamformer, which involves the inverse of the sample covariance matrix, degrades significantly for single-snapshot data and correlated sources, Figs. 1(b) and 1(c). Regularization of the MVDR weights Eq. (9), here $\beta = \sigma^2$, smooths the MVDR estimate towards the low-resolution CBF estimate. The sparsity promoting SBL beamformer offers high-resolution reconstruction, with single-snapshot data and correlated arrivals invariably, even at low SNR Fig. 1(d). The spurious peaks (e.g., around −45°) at the SBL power spectrum, $\widehat{\boldsymbol{\xi}}$, for low SNR, Fig. 1(d), do not affect the unbiased amplitude estimate, $\widehat{\mathbf{X}}_{\mathrm{SBL},\mathcal{N}} = \mathbf{A}_{\mathcal{N}}^{+}\mathbf{Y}$, as

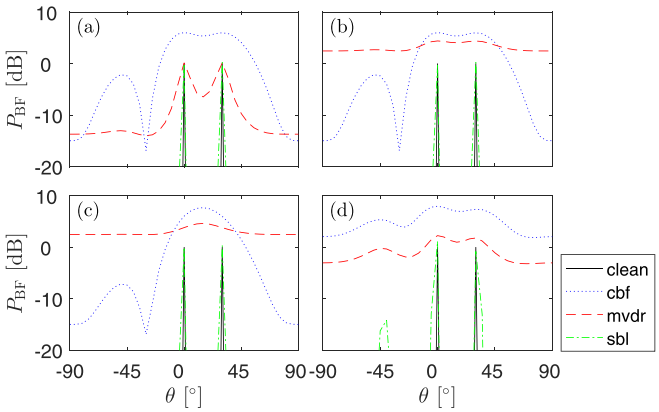

FIG. 1. (Color online) CBF, MVDR, and SBL power spectra from L snapshots for two equal-strength sources at 0° and 30° as the clean signal with a uniform linear array with $M = 4$ sensors and spacing $d/\lambda = 1/2$ for (a) SNR = 20 dB, $L = 2M$, uncorrelated sources, (b) SNR = 20 dB, $L = 1$, uncorrelated sources, (c) SNR = 20 dB, $L = 2M$, correlated sources, (d) SNR = 0 dB, $L = 2M$, uncorrelated sources.

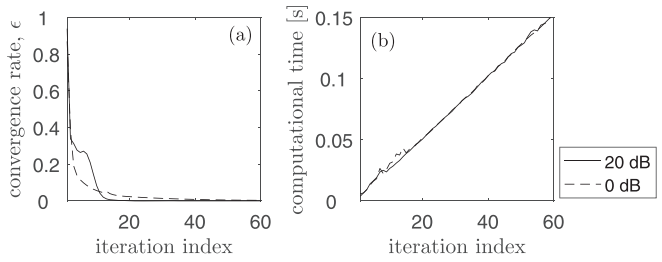

FIG. 2. (a) Convergence rate ϵ and (b) computational time of the SBL beamformer algorithm per number of iterations, at SNR = 20 dB (solid line) and 0 dB (dashed line).

the sparsity level is set to $K = 2$ at step 5 of the algorithm in Table I.

The results in Fig. 1 indicate that the SBL beamformer offers robust DOA estimation, particularly in case of snapshot-starved data, e.g., for non-stationary signals, and reverberant environments. Opposed to the CBF and MVDR beamformers which are implemented as spatial filters, the SBL beamformer involves an iterative estimation of the likelihood and prior hyperparameters. Figure 2 shows that the convergence rate, ε, decreases rapidly with the number of iterations while the CPU time on an Intel Core i5 increases linearly. The computational time for a single SBL iteration is ca. 3 ms compared to ca. 0.07 ms for CBF and ca. 0.1 ms for MVDR. Nevertheless, the reconstruction accuracy of SBL is significant. Notably, the computational time per number of snapshots is almost constant.[26]

In the following, the parameters of the SBL algorithm in Table I are set to $N_{\text{iter}} = 20$ and $\epsilon_{\text{min}} = 0.001$. These values offer adequate estimation accuracy and computational efficiency (see Fig. 2) for problems of small dimensions, e.g., $M = 4$, $N = 37$, which are typical[31,39] for the speech processing applications in focus. More iterations might be required for the SBL algorithm to converge for larger problems.[26] The sparsity K is set to the number of sources in each case. Since speech is broadband, the multi-frequency update rule (30) is used for the SBL reconstruction.

IV. SIMULATION RESULTS

A listening scenario of interest where speech enhancement and separation is beneficial for speech intelligibility involves focusing at a reference talker in the presence of noise, competing talkers and reverberation. The performance of CBF, MVDR, and SBL beamforming in such conditions is demonstrated, herein, with simulations.

For the simulations, a ULA (31) is considered with $M = 4$ sensors. The inter-sensor spacing is $d = 28.6$ cm to avoid spatial aliasing, i.e., $d/\lambda < 1/2$, for frequencies up to 6 kHz which is the upper frequency for high speech quality, assuming airborne propagation with sound speed $c = 343$ m/s. The sources are speech excerpts from the EUROM1 English corpus[40] including both male and female talkers of 1 s duration resampled at $f_s = 16$ kHz. The speech excerpts, due to their short duration, have constant voice activity without silent intervals. Hence, the root-mean-square value of the target source, $\text{rms}(x_{\text{target}}) = \sqrt{1/T \sum_{t=1}^{T} x_{\text{target}}(t)^2}$ where T is the total number of samples, is used to determine the noise variance in relation to the SNR, $\sigma^2 = 10^{-\text{SNR}/10}\text{rms}(x_{\text{target}})^2$. A DOA grid $[-90°: 5°: 90°]$ is considered.

The signals are processed in 40 ms frames with 10% overlap. Each frame is further divided in 8 ms snapshots with 50% overlap resulting in $L = 9$ snapshots per frame. This way, the signal per frame can be approximated as stationary while having enough snapshots $L > 2M$ for a statistically robust sample data cross-spectral matrix $\mathbf{S_y}$ (as in Ref. 41). A Hanning window is applied to each snapshot followed by a STFT. The resulting narrowband signals, for each frequency in the resulting spectrum ranging 0−8 kHz, are processed with steered beamforming methods for DOA estimation as detailed in Sec. III. Finally, for each direction on the resulting DOA map, an inverse STFT is applied to the reconstructed signals which are resynthesized to the time domain with the overlap-and-add procedure.[42]

Figure 3 depicts the DOA maps for the simple case of a single talker in the presence of additive noise at SNR = 15 dB, along with the spectrograms of the reconstructed signals at selected directions, calculated over frames of 40 ms duration, Hanning weighting and 50% overlap. Specifically, Fig. 3(a) indicates the actual source distribution across time and DOA. There is a single source of male speech at $\theta = 50°$ with frequency spectrum per time frame shown in the spectrogram in Fig. 3(b). The CBF, MVDR, and SBL estimates are depicted in Figs. 3(c), 3(f), and 3(i), respectively. In this case with a single source, additive noise at high SNR and sufficient snapshots, all methods

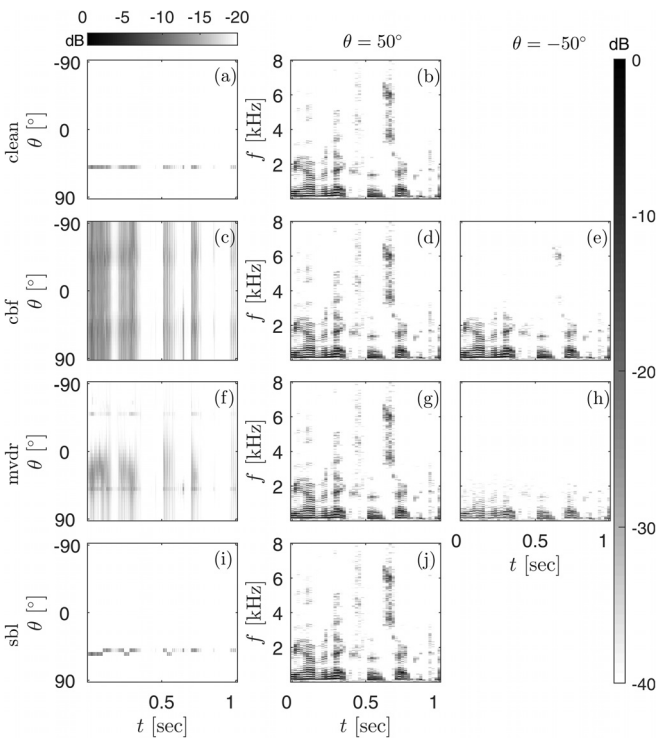

FIG. 3. DOA maps for a single source (male talker) at 50° with additive noise at SNR = 15 dB for (a) the original signal, (c) CBF, (f) MVDR, and (i) SBL reconstruction. Spectrograms of the (b) clean signal, (d) CBF, (g) MVDR, and (j) SBL estimates at $\theta = 50°$. Spectrograms of the (e) CBF and (h) MVDR estimates at $\theta = -50°$.

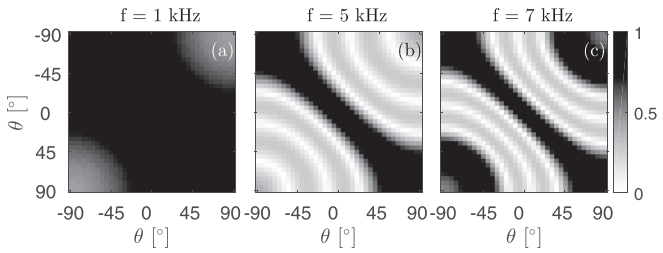

FIG. 4. Gram matrices $1/M(\mathbf{A}^H\,\mathbf{A})$ indicating the coherence pattern of the steering vectors (31) for a ULA with $M=4$ sensors and $d=28.6$ cm uniform spacing at (a) $f=1$ kHz, (b) $f=5$ kHz and (c) $f=7$ kHz.

reconstruct accurately the target signal at $\theta=50°$ as shown in the corresponding spectrograms, Figs. 3(d), 3(g), and 3(j).

However, the low resolution CBF spreads the energy across the whole angular spectrum making DOA estimation very difficult. For example, there is a lot of energy at $\theta=-50°$, especially at low frequencies, due to the single source at $\theta=50°$; see Fig. 3(e). This is explained by the coherence of the steering vectors (31) at different frequencies as indicated by the Gram matrices, $1/M(\mathbf{A}^H\,\mathbf{A})$, in Fig. 4. Note that each row of the Gram matrix, $1/M[\mathbf{a}^H(\theta)\,\mathbf{A}]$, is the CBF beampattern for a unit source at θ. At low frequencies the array aperture is too small to detect phase differences of the recorded wavefield across sensors and the CBF estimate is almost omnidirectional, Fig. 4(a). The CBF

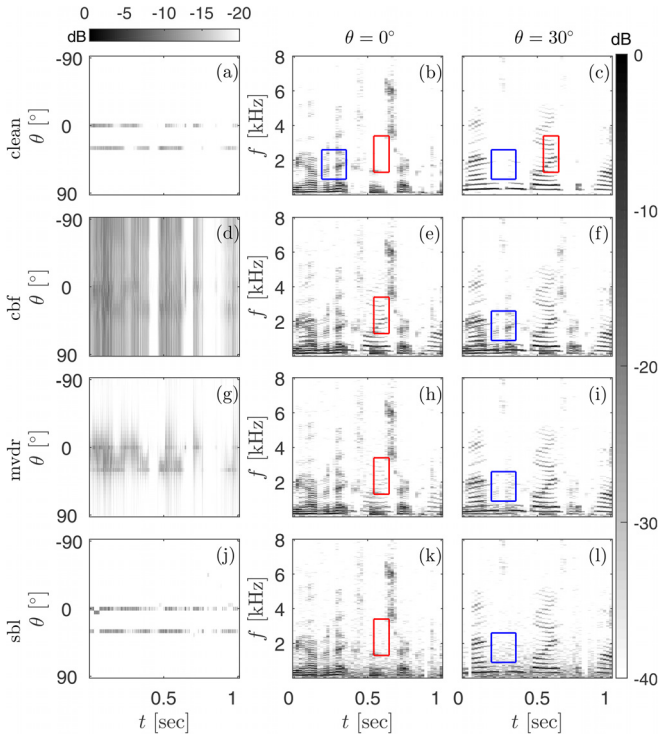

FIG. 5. (Color online) DOA maps for a source (male talker) at 0° and a source (female talker) at 30° with additive noise at SNR = 15 dB for (a) the original signal, (d) CBF, (g) MVDR, and (j) SBL reconstruction. Spectrograms of the (b) clean signal, (e) CBF, (h) MVDR, and (k) SBL estimates at $\theta=0°$. Spectrograms of the (c) clean signal, (f) CBF, (i) MVDR, and (l) SBL estimates at $\theta=30°$. The blue box indicates an example of a time-frequency region where there is significant energy from the source at 0° and almost no energy from the source at 30° and vice versa within the red box.

estimate becomes more directive for higher frequencies, Fig. 4(b), while for $d/\lambda>1/2$ grating lobes appear in the estimate due to spatial aliasing Fig. 4(c). The directionality characteristics of CBF depicted in Fig. 4 indicate that processing only higher frequencies (e.g., above 2 kHz for the particular configuration) could improve the corresponding DOA estimates. However, this is not a suitable option for short-time processing of speech signals which have only a few energy (if any) at high frequencies as DOA estimation would fail due to absence of signal.

MVDR improves the resolution, Fig. 3(h), while SBL offers very accurate DOA estimation. Note that the spectrograms for the signal at $\theta=-50°$ in the clean and SBL DOA map are omitted since their energy is below the plotted dynamic range.

Figure 5 demonstrates the DOA estimation performance of CBF, MVDR, and SBL beamforming in the case of two sources, namely, a male talker at 0° and a female talker at 30° as shown in Fig. 5(a), and additive noise at SNR = 15 dB. The low resolution CBF offers smooth DOA reconstruction, Fig. 5(d), which results in poor localization hence poor signal separation. For example, the CBF estimate at 0° [Fig. 5(e)] contains energy not only from the source at 0° [Fig. 5(b)] but also from the source at 30° [Fig. 5(c)] and vice versa [Fig. 5(f)]. The MVDR estimate has improved resolution [Fig. 5(g)], attenuating more effectively signals from directions other than the focusing one [Figs. 5(h) and 5(i)]. SBL offers great spatial selectivity hence source separation [Figs. 5(j)–5(l)].

Finally, Fig. 6 shows the corresponding results to Fig. 5 when the source at 30° is a replica of the source at 0°. In this case, the sources are correlated, e.g., in the presence of strong reflections due to reverberant listening

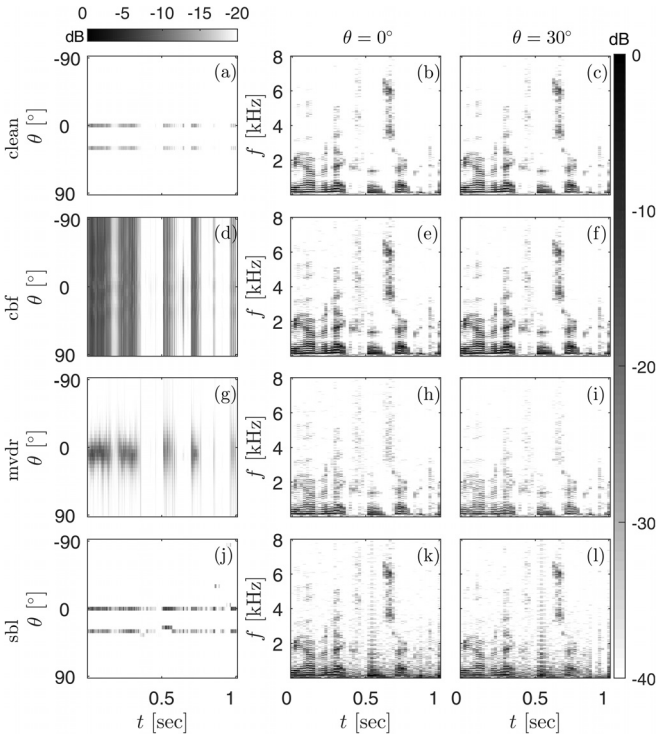

FIG. 6. The respective DOA maps and spectrograms as in Fig. 5 replacing the signal at 30° with a replica of the signal at 0°.

environments, and the MVDR estimate degenerates, merging the two sources into one and localizing it in between the true source directions [Figs. 6(g)–6(i)]. The SBL beamformer, localizes the two coherent sources accurately [Figs. 6(j)–6(l)].

A. Performance metrics

The results in Figs. 3 and 5 and 6 indicate qualitatively the performance of CBF, MVDR, and SBL DOA estimation in the presence of both uncorrelated and correlated sources under high-SNR listening conditions. To evaluate the performance of CBF, MVDR, and SBL beamforming quantitatively as a function of SNR, the following performance metrics are introduced:

(1) The relative root-mean-square error at the focusing direction,

$$\mathrm{rrmse}_{\theta_f} = \sqrt{\frac{\frac{1}{N_t}\sum_{t\in T}|x(\theta_f,t)-\widehat{x}(\theta_f,t)|^2}{\frac{1}{N_t}\sum_{t\in T}|x(\theta_f,t)|^2}}, \tag{32}$$

which indicates the relative noise level of the reconstructed signal $\widehat{x}(\theta_f,t)$ at the focusing direction θ_f with respect to the clean signal $x(\theta_f,\ t)$, such that $\widehat{\mathrm{SNR}}_{\theta_f} = -20\log_{10}(\mathrm{rrmse}_{\theta_f})$ dB. The rrmse for the unprocessed data, e.g., the recorded signal at the mth microphome $y_m(t)$, indicates the relative noise level in the measurements, yielding the SNR. Hence, the SNR improvement due to the beamforming estimate is $(\widehat{\mathrm{SNR}}_{\theta_f} - \mathrm{SNR})$ dB.

(2) The beamformer's directivity

$$\mathrm{D} = \frac{\frac{1}{N_t}\sum_{t\in T}|\widehat{x}(\theta_f,t)|^2}{\frac{1}{N_t}\sum_{t\in T}\frac{1}{N}\sum_{\theta\in\Theta}|\widehat{x}(\theta,t)|^2}, \tag{33}$$

or equivalently the directivity index $\mathrm{DI} = 10\log_{10}\mathrm{D}$ dB, which indicates the ratio of the power of the reconstructed signal $\widehat{x}(\theta_f,t)$ at the focusing direction θ_f to the mean power of the reconstructed signal $\widehat{x}(\theta,t)$ over all N directions on the angular grid. Thus, for an omnidirectional signal $x_{\mathrm{omni}}(\theta_f,t) = x_{\mathrm{omni}}(\theta,t)$, $\forall\,\theta\in\Theta$, i.e., the mean power over all directions on the grid is equal to the power at the focusing direction and $\mathrm{D}=1$ or $\mathrm{DI}=0$ dB. The more a beamformer suppresses the signal from directions other the focusing one, the larger is its directivity and the more accurate the DOA estimate. For a superdirective beamformer, such that $\widehat{x}(\theta,t)=0$, $\{\theta:\theta\in\Theta|\theta\neq\theta_f\}$, the directivity is maximized, $\mathrm{D}=N$.

(3) The short-time objective intelligibility (STOI) measure[43] which is used to predict the speech intelligibility of the beamformed signal, hence evaluate perceptual consequences of the beamforming algorithm. STOI receives as inputs a clean reference signal and a degraded version of it due to noise and/or distortion and outputs the correlation coefficient (0 for unintelligible speech, 1 for fully intelligible speech) between the temporal envelopes of the input signals in short-time (384 ms) segments. STOI correlates well with subjective evaluation of speech intelligibility, i.e., from listening experiments.

The performance of CBF, MVDR, and SBL beamforming in reconstructing a target source at 0° in the presence of additive noise at a range of [−5:5:15] dB SNR is evaluated. Two noise types are examined, broadband white noise and babble noise constructed by overlapping speech from six talkers in the EUROM1 English corpus.[40] For each noise type and at each SNR, beamforming estimates are obtained for 100 random realizations of speech and noise. The mean statistics of the performance metrics, namely, the rrmse at the focusing direction (32), the directivity (33), and the STOI score, are shown in Fig. 7.

All beamforming methods improve the SNR when focused at the direction of the target source compared to the SNR of the omnidirectional data for both noise types, Fig. 7(a). Consequently, the speech signal at 0° is enhanced over noise as indicated by the STOI scores in Fig. 7(c). However, the conventional CBF and MVDR beamformers have low directivity, Fig. 7(b), resulting in low resolution DOA maps with energy across the whole angular spectrum; e.g., see Figs. 3(c) and 3(f). Only the superdirective SBL beamformer, Fig. 7(b), offers unambiguous DOA estimation.

V. EXPERIMENTAL RESULTS

The high-resolution DOA estimation and speech separation capabilities of SBL are validated with experimental data in multi-talker, noisy, reverberant listening conditions. The measurement prototype comprises a workshop safety helmet circularly perforated above the cap and 8 microphones, which are adjusted on the front part of the helmet on a semicircular configuration with a uniform angular spacing 22.5°. The sensing matrix $\mathbf{A}$ for this array configuration is determined experimentally through the HRTFs. To obtain the HRTFs, the helmet is fitted on a Knowles electronics mannequin for acoustics research (KEMAR) and placed on a turning-base in the anechoic chamber at GN Hearing A/S, Ballerup, Denmark. Impulse responses are recorded for all microphones at a sampling frequency $f_s = 24\,414$ Hz, sequentially while rotating KEMAR by 2°

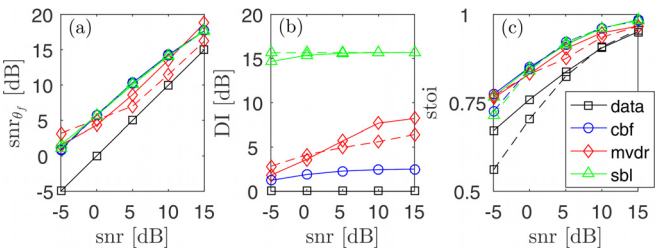

FIG. 7. (Color online) Mean values of (a) the $\widehat{\mathrm{SNR}}_{\theta_f}$ at $\theta_f=0°$, (b) the directivity index DI, and (c) the STOI score for CBF, MVDR, and SBL beamforming reconstruction of a target source at 0° in the presence of white (solid lines) or babble noise (dashed lines) as a function of SNR from 100 random realizations. For comparison, the corresponding values for the data, i.e., the unprocessed signal from the first microphone on the array, are depicted.

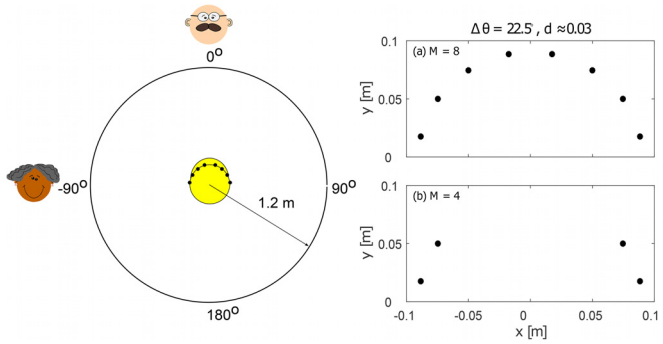

FIG. 8. (Color online) Measurement setup and microphone positions for the considered array configurations.

until completing a full-circle rotation ($\theta = 0°$: $2°$: $360°$, $N = 181$).

The measurement setup involves two speakers, the first exactly in front of KEMAR, at $0°$, playing 2 s of male speech and the other towards the left ear, at $-90°$, playing 2 s of female speech. Both speakers were elevated to the plane of the array and placed at a radial distance of 1 m from KEMAR; see Fig. 8. The arrangement is set in an anechoic chamber and measurements are taken considering the full array as shown in Fig. 8(a) as a reference scenario, as well as in a populated canteen considering only the four microphones that are lying above the ears as shown in Fig. 8(b), as a challenging listening environment. All locations are at the facilities of GN Hearing A/S, Ballerup, Denmark. The signals are processed in single-snapshot, 20 ms frames with 50% overlap. A Hanning window followed by a STFT is applied to each frame and the resulting narrowband signals are beamformed with CBF and SBL. MVDR beamforming is omitted here due to the single-snapshot processing. The resulting steered responses are resynthesized with the overlap-and-add procedure.[42]

Figure 9 shows the DOA maps of the clean and the recorded signal in anechoic conditions and the CBF and SBL DOA estimates along with the corresponding spectrograms (calculated over frames of 40 ms duration, Hanning weighting and 50% overlap) at the speaker locations, i.e., at $0°$ and $-90°$, respectively. The two speech signals, Figs. 9(b) and 9(c), are mixed in the unprocessed single-microphone recording, Figs. 9(e) and 9(f), which does not offer directional information, Fig. 9(d). CBF attributes directivity to the microphone array by attenuating wavefield contributions from directions other than the focusing one, Figs. 9(h) and 9(i), but has low resolution, Fig. 9(g). The high-resolution SBL beamformer not only localizes accurately the two speakers, Fig. 9(j), but also separates the corresponding speech signals, Figs. 9(k) and 9(l), validating the simulation results, e.g., compare with Fig. 5. Similarly, Fig. 10, demonstrates the corresponding results for measurements in a populated canteen with reverberation time $T_{60} = 0.9$ s, at SNR $= -6$ dB. In this case, the recorded signal is very noisy due to babble, clinking cutlery, reverberation, etc., thus, both CBF and SBL DOA estimates deteriorate accordingly. Nevertheless, the SBL beamformer suppresses noise more effectively.

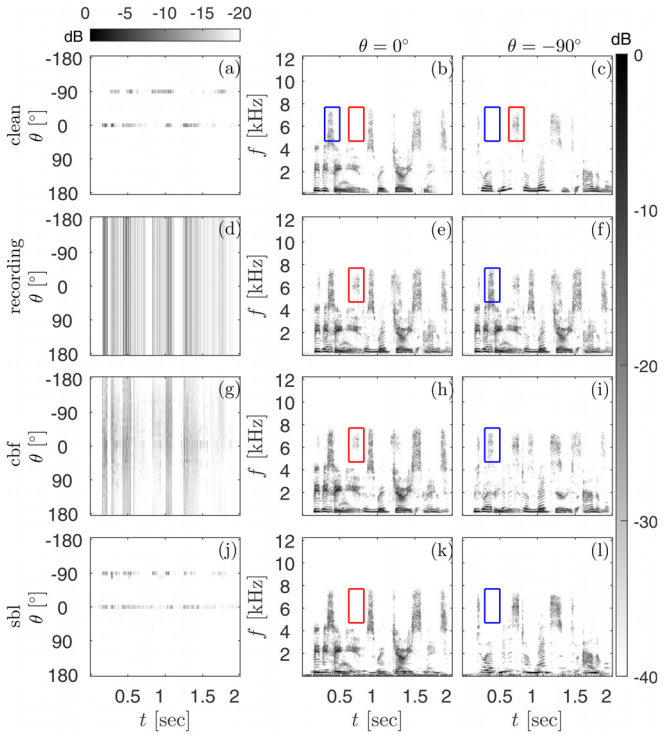

FIG. 9. (Color online) DOA maps obtained with the array configuration in Fig. 8(a) for a source (male talker) at $0°$ and a source (female talker) at $-90°$ in anechoic conditions for (a) the original signals, (d) the recorded signal from the front left microphone, (g) CBF, and (j) SBL reconstruction. Spectrograms of the (b) clean signal, (e) recorded signal, (h) CBF, and (k) SBL estimates at $\theta = 0°$. Spectrograms of the (c) clean signal, (f) recorded signal, (i) CBF, and (l) SBL estimates at $\theta = -90°$. The blue box indicates an example of a time-frequency region where there is significant energy from the source at $0°$ and almost no energy from the source at $-90°$ and vice versa within the red box.

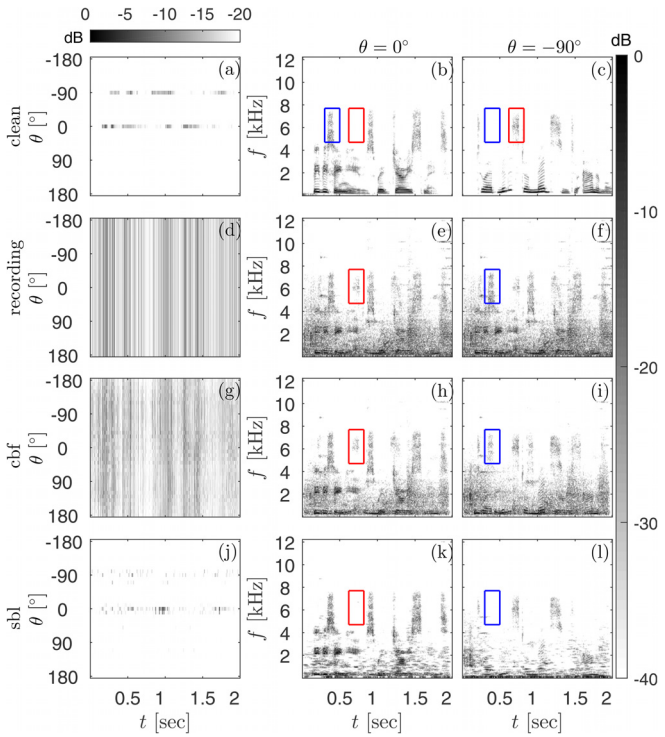

FIG. 10. (Color online) The respective DOA maps and spectrograms as in Fig. 9 for signals recorded with the array configuration in Fig. 8(b) in a canteen.

VI. CONCLUSION

We use a probabilistic sparse signal reconstruction approach to solve simultaneously the sound source localization and speech enhancement problem within the SBL framework. The SBL formulation offers sparse robust DOA estimates by auto-regularizing a hierarchical Bayesian model with adaptive selection of the hyperparameters from the data.

Contrary to established spatial filtering methods, SBL beamforming provides high-resolution acoustic imaging even with correlated arrivals and single-snapshot measurements. Both simulation results with a ULA and experimental measurements with a semi-circular prototype array show that SBL beamforming offers simultaneous sound source localization and separation offering speech enhancement over noise, reverberation and competing talkers.

ACKNOWLEDGMENTS

This work was supported by the Innovation Fund Denmark, under Grant No. 99-2014-1.

[1]M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *IEEE Int. Conf. on Acoust., Speech, and Signal Process (ICASSP-12)* (IEEE, New York, 2012), pp. 409–412.
[2]J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Statistically efficient methods for pitch and DOA estimation," in *IEEE Int. Conf. on Acoust., Speech, and Signal Process. (ICASSP-13)* (IEEE, New York, 2013), pp. 3900–3904.
[3]J. A. MacDonald, "A localization algorithm based on head-related transfer functions," J. Acoust. Soc. Am. **123**, 4290–4296 (2008).
[4]F. Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," IEEE Trans. Instrum. Meas. **63**, 2098–2107 (2014).
[5]C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," IEEE Trans. Multimedia **10**, 538–548 (2008).
[6]M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," IEEE/ACM Trans. Audio, Speech, and Lang. Proc. **25**, 611–623 (2017).
[7]R. Giri, B. D. Rao, F. Mustiere, and T. Zhang, "Dynamic relative impulse response estimation using structured sparse Bayesian learning," in *IEEE Int. Conf. on Acoust., Speech, and Signal Process. (ICASSP-16)* (IEEE, New York, 2016), pp. 514–518.
[8]H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," IEEE Signal Proc. Mag. **13**, 67–94 (1996).
[9]G. F. Edelmann and C. F. Gaumond, "Beamforming using compressive sensing," J. Acoust. Soc. Am. **130**, 232–237 (2011).
[10]S. Fortunati, R. Grasso, F. Gini, M. S. Greco, and K. LePage, "Single-snapshot DOA estimation by using compressed sensing," EURASIP J. Adv. Signal Process. **120**, 1–17 (2014).
[11]A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," J. Acoust. Soc. Am. **136**, 260–271 (2014).
[12]P. Gerstoft, A. Xenaki, and C. Mecklenbräuker, "Multiple and single snapshot compressive beamforming," J. Acoust. Soc. Am. **138**, 2003–2014 (2015).
[13]Z. Tang, G. Blacquière, and G. Leus, "Aliasing-free wideband beamforming using sparse signal representation," IEEE Trans. Signal Proc. **59**, 3464–3469 (2011).
[14]G. Chardon, L. Daudet, A. Peillot, F. Ollivier, N. Bertin, and R. Gribonval, "Near-field acoustic holography using sparse regularization and compressive sampling principles," J. Acoust. Soc. Am. **132**, 1521–1534 (2012).
[15]E. Fernandez-Grande, A. Xenaki, and P. Gerstoft, "A sparse equivalent source method for near-field acoustic holography," J. Acoust. Soc. Am. **141**, 532–542 (2017).
[16]A. Xenaki, E. Fernandez-Grande, and P. Gerstoft, "Block-sparse beamforming for spatially extended sources in a Bayesian formulation," J. Acoust. Soc. Am. **140**, 1828–1838 (2016).
[17]M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," J. Mach. Learn. Res. **1**, 211–244 (2001).
[18]D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," IEEE Trans. Signal Proc. **52**, 2153–2164 (2004).
[19]M. A. T. Figueiredo, "Adaptive sparseness for supervised learning," IEEE Trans. Pattern Anal. Mach. Intell. **25**, 1150–1159 (2003).
[20]S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian compressive sensing using Laplace priors," IEEE Trans. Image Process. **19**, 53–63 (2010).
[21]R. Giri and B. D. Rao, "Type I and Type II Bayesian methods for sparse signal recovery using scale mixtures," IEEE Trans. Signal Process. **64**, 3418–3428 (2016).
[22]D. P. Wipf and S. Nagarajan, "Beamforming using the relevance vector machine," in *Int. Conf. on Machine Learning* (ACM, New York, 2007), pp. 1023–1030.
[23]Z.-M. Liu, Z.-T. Huang, and Y.-Y. Zhou, "An efficient maximum likelihood method for direction-of-arrival estimation via sparse Bayesian learning," IEEE Trans. Wireless Commun. **11**, 1–11 (2012).
[24]E. Zhang, J. Antoni, B. Dong, and H. Snoussi, "Bayesian space-frequency separation of wide-band sound sources by a hierarchical approach," J. Acoust. Soc. Am. **132**, 3240–3250 (2012).
[25]A. Pereira, J. Antoni, and Q. Leclere, "Empirical Bayesian regularization of the inverse acoustic problem," Appl. Acoust. **97**, 11–29 (2015).
[26]P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multisnapshot sparse Bayesian learning for DOA," IEEE Signal Process. Lett. **23**, 1469–1473 (2016).
[27]K. L. Gemba, S. Nannuru, P. Gerstoft, and W. S. Hodgkiss, "Multi-frequency sparse Bayesian learning for robust matched field processing," J. Acoust. Soc. Am. **141**, 3411–3420 (2017).
[28]S. Nannuru, K. L. Gemba, P. Gerstoft, W. S. Hodgkiss, and C. F. Mecklenbräuker, "Sparse Bayesian learning with uncertainty models and multiple dictionaries," arXiv:1704.00436v2 (2017).
[29]X. Wu, W.-P. Zhu, and J. Yan, "Direction of arrival estimation for off-grid signals based on sparse Bayesian learning," IEEE J. Sens. **16**, 2004–2016 (2016).
[30]J. Meyer, "Beamforming for a circular microphone array mounted on spherically shaped objects," J. Acoust. Soc. Am. **109**, 185–193 (2001).
[31]D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov, "Fast head-related transfer function measurement via reciprocity," J. Acoust. Soc. Am. **120**, 2202–2215 (2006).
[32]H. Van Trees, *Optimum Array Processing* (Wiley-Interscience, New York, 2002), Chaps. 1–10.
[33]J. Capon, "High-resolution frequency-wavenumber spectrum analysis," Proc. IEEE **57**, 1408–1418 (1969).
[34]S. Nadarajah, "A generalized normal distribution," J. Appl. Stat. **32**, 685–694 (2005).
[35]Z. He, S. Xie, S. Ding, and A. Cichocki, "Convolutive blind source separation in the frequency domain based on sparse representation," IEEE Trans. Audio, Speech, Lang. Proc. **15**, 1551–1563 (2007).
[36]R. Tibshirani, "Regression shrinkage and selection via the lasso," J. R. Stat. Soc. Ser. B **58**, 267–288 (1996).
[37]S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, New York, 2004), pp. 1–684.
[38]A. Koochakzadeh and P. Pal, "On saturation of the Cramér Rao bound for sparse Bayesian learning," in *IEEE Int. Conf. on Acoust., Speech, and Sig. Proc. (ICASSP-17)* (IEEE, New York, 2017), pp. 3081–3085.
[39]H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural impulse responses," EURASIP J. Adv. Signal Proc. **2009**, 1–10 (2009).
[40]D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and, J. Zieliger, "EUROM-a spoken language resource for the EU," in *Europ. Conf. on Speech Commun. and Speech Tech. (Eurospeech-95)* (1995), Vol. 1, pp. 867–880.
[41]A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Europ. Signal Process. Conf. (EUSIPCO-14)* (IEEE, New York, 2014), pp. 61–65.
[42]T. F. Quatieri, "Discrete-time speech signal processing: Principles and practice," in *Signal Processing* (Pearson Education India, NJ, 2006), Chap. 7.
[43]C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Proc. **19**, 2125–2136 (2011).